Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at http://about.jstor.org/participate-jstor/individuals/early-journal-content.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

SCIENCE.—SUPPLEMENT.

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

AUGUSTUS DEMORGAN somewhere remarks (I think it is in his 'Budget of paradoxes') that some time somebody will institute a comparison among writers in regard to the average length of words used in composition, and that it may be found possible to identify the author of a book, a poem, or a play, in this way.

In reflecting upon this remark at various times within the past five or six years, always with the determination to test the value of the suggestion whenever time for the work seemed available, a more comprehensive and satisfactory method of analysis than that based simply upon mean word-length suggested itself. The new method, while scarcely more laborious than that proposed by DeMorgan, promised to yield results more quickly and of a definitely higher order. It also had the advantage of including, in its application, all that was necessary to the determination of mean word-length; so that, in reality, it furnished two distinct tests.

Preliminary trials of the method have furnished strong grounds for the belief that it may prove useful as a method of analysis leading to identification or discrimination of authorship, and it is therefore brought to the attention of the scientific and literary public in the hope that some one may be found who is at once able and willing to secure a satisfactory test of its validity.

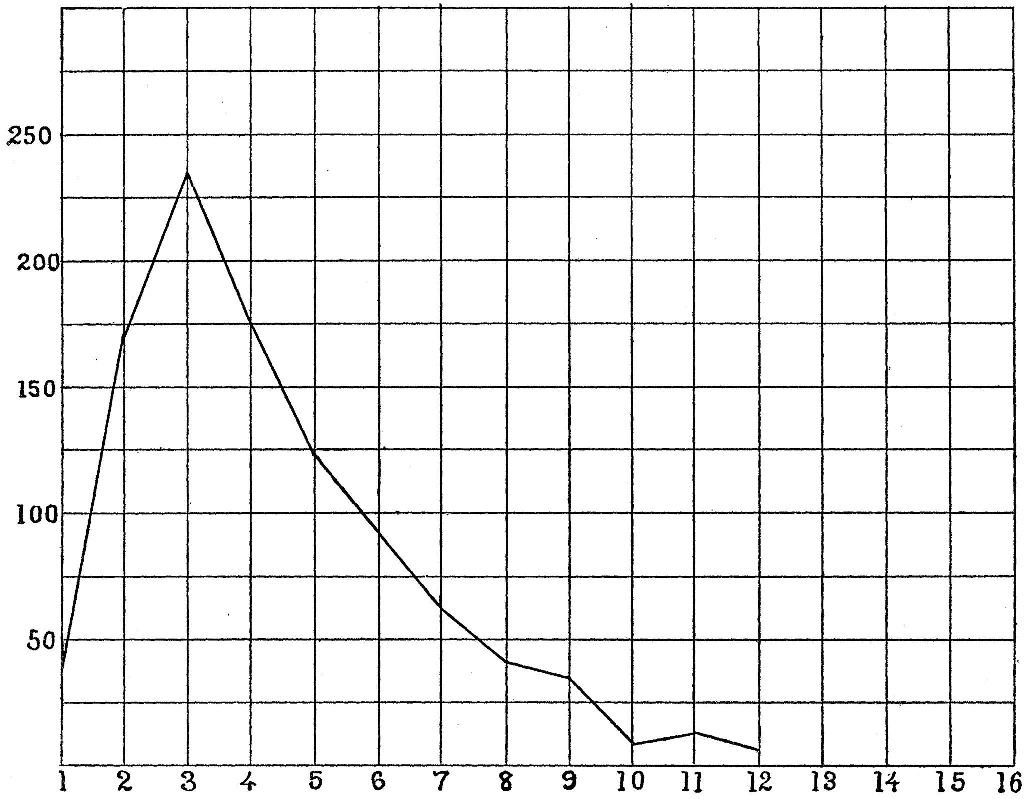

FIG. 1. — FIRST ONE THOUSAND WORDS IN 'OLIVER TWIST.'

The nature of the process is extremely simple, but it may be useful to point out its similarity to a well-known method of material analysis, the consideration of which actually first suggested to the writer its literary analogue.

By the use of the spectroscope, a beam of non-homogeneous light is analyzed, and its components assorted according to their wave-length. As is well known, each element, when intensely heated under proper conditions, sends forth light which, upon prismatic analysis, is found to consist of groups of waves of definite length, and appearing in certain definite proportions. So certain and uniform are the results of this analysis, that the appearance of a particular spectrum is indisputable evidence of the presence of the element to which it belongs.

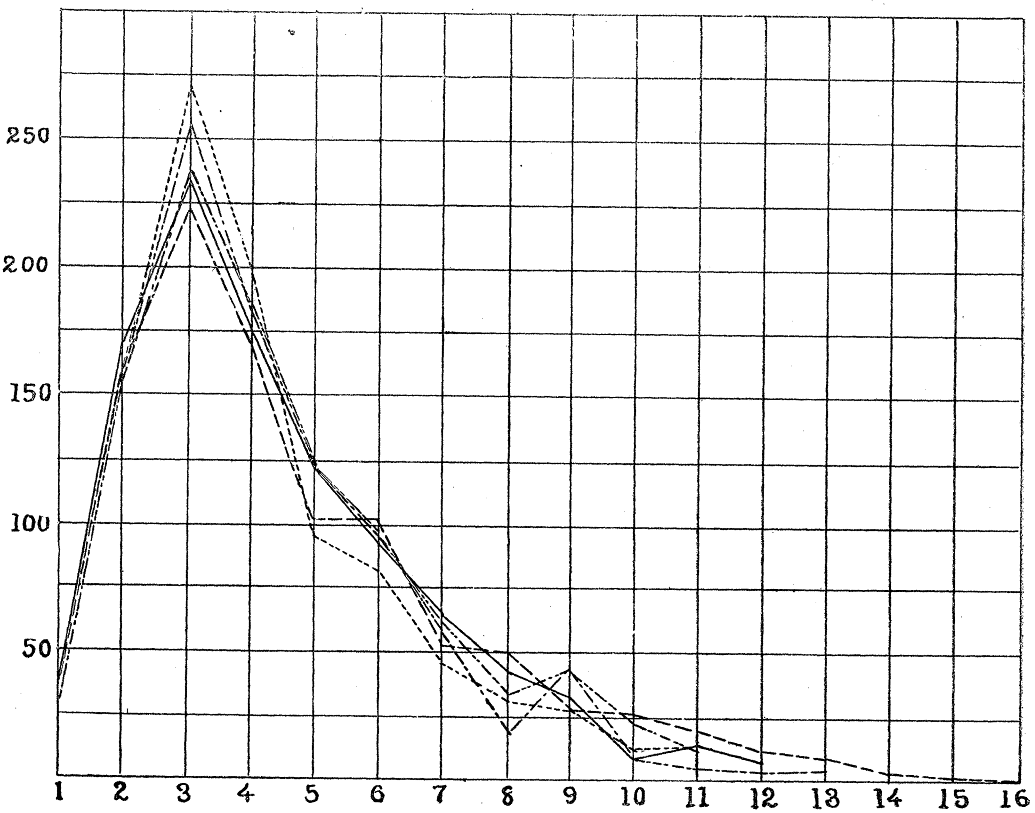

Fig. 2. — Showing five groups, of one thousand words each, from 'Oliver Twist.'

In a manner very similar, it is proposed to analyze a composition by forming what may be called a 'word-spectrum,' or 'characteristic curve,' which shall be a graphic representation of an arrangement of words according to their length and to the relative frequency of their occurrence. If, now, it shall be found that with every author, as with every element, this spectrum persists in its form and appearance, the value of the method will be at once conceded. It has been proved that the spectrum of hydrogen is the same, whether that element is obtained from the water of the ocean or from the vapor of the atmosphere. Wherever and whenever it appears, it means hydrogen. If it can be proved that the word-spectrum or characteristic curve exhibited by an analysis of 'David Copperfield' is identical with that of 'Oliver Twist,' of 'Barnaby Rudge,' of 'Great expectations,' of the 'Child's history of England,' etc., and that it differs sensibly from that of 'Vanity fair,' or 'Eugene Aram,' or 'Robinson Crusoe,' or 'Don Quixote,' or any thing else in fact, then the conclusion will be tolerably certain that when it appears it means Dickens.

The validity of the method as a test of authorship, then, implies the following assumptions: that every writer makes use of a vocabulary which is peculiar to himself, and the character of which does not materially change from year to year during his productive period; that, in the use of that vocabulary in composition, personal peculiarities in the construction of sentences will, *in the long-run*, recur with such regularity that short words, long words, and words of medium length, will occur with definite relative frequencies.

The first assumption will, perhaps, be admitted in a general way, without debate. It is easily seen that to prove or disprove the second will require the expenditure of an enormous amount of labor. The following results are offered as a means of properly exhibiting the method, and as evidence, in some degree at least, of its real value.

It is important, first, to determine to what extent an author may be said to agree with himself; and, second, to what extent does he differ from others.

As an instance in which two writers might well be expected to greatly resemble each other in their curves, and consequently as a severe test of the method, two contemporaneous novelists, Dickens and Thackeray, were selected for the first examination. The operation consisted simply in counting the number of letters in every word, and recording the number of words of one letter, two letters, three letters, etc. The count began in both cases at the beginning of the volume, and, after a few thousand words had been counted in order, the book was opened at random near the middle, and the count continued. In no case was any personal choice exercised, except that both counts began with the first chapter. Words were counted always in groups of one thousand. The graphic display of the result was made by the common method of rectangular co-ordinates, using the number of letters in a word as an abscissa, and the corresponding number of such words in a thousand as an ordinate. As an illustration, the first one thousand words counted from 'Oliver Twist' may be cited; they were as follows: —

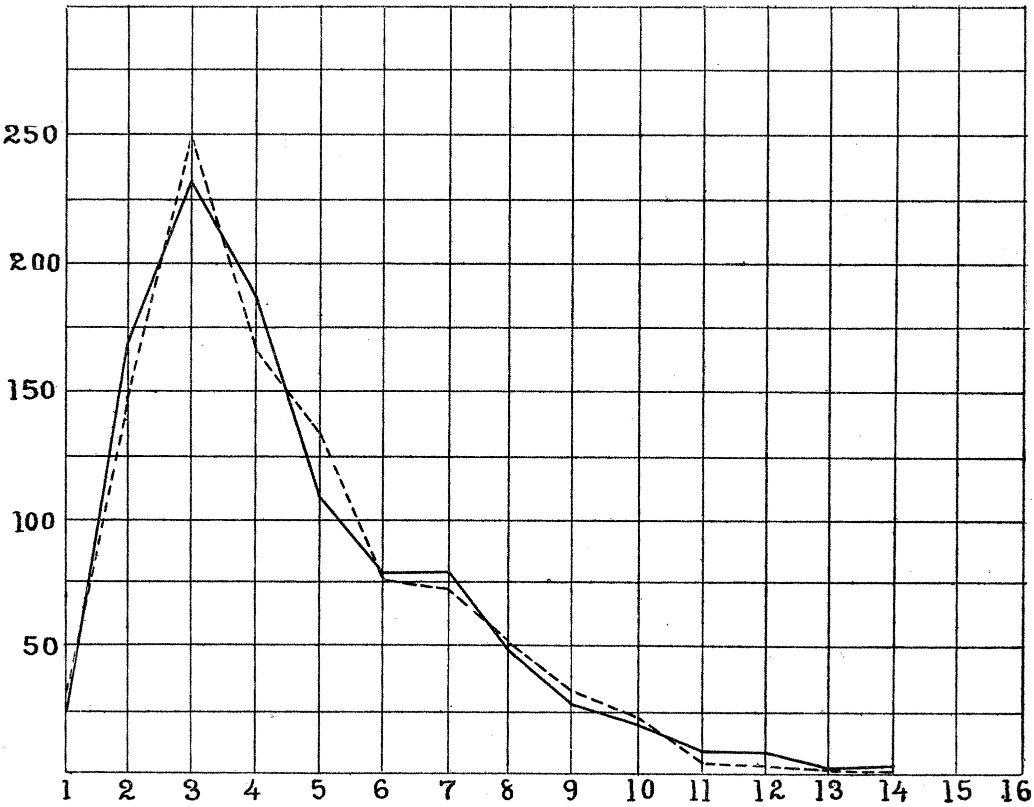

Fig. 3. — Two consecutive groups, of one thousand words each, from 'Vanity Fair.' These groups show sensibly the same average word-lengths.

Number of letters	1	2	3	4	5	6	7	8	9	10	11	12
Number of words	38	170	235	175	123	91	62	41	35	10	13	7

Even in so small a number as one thousand, the relative distribution of words is approximately the same as in a much larger number, although, as would naturally be expected, accidental variations or 'runs' overshadow personal characteristics to a great extent; but not completely, as will be seen in the characteristic curves shown in the following pages. In fact, when the ten groups, of a thousand words each, from Dickens, are compared with ten similar groups from John Stuart Mill, no one of the first set could by any possibility be mistaken for any one of the second.

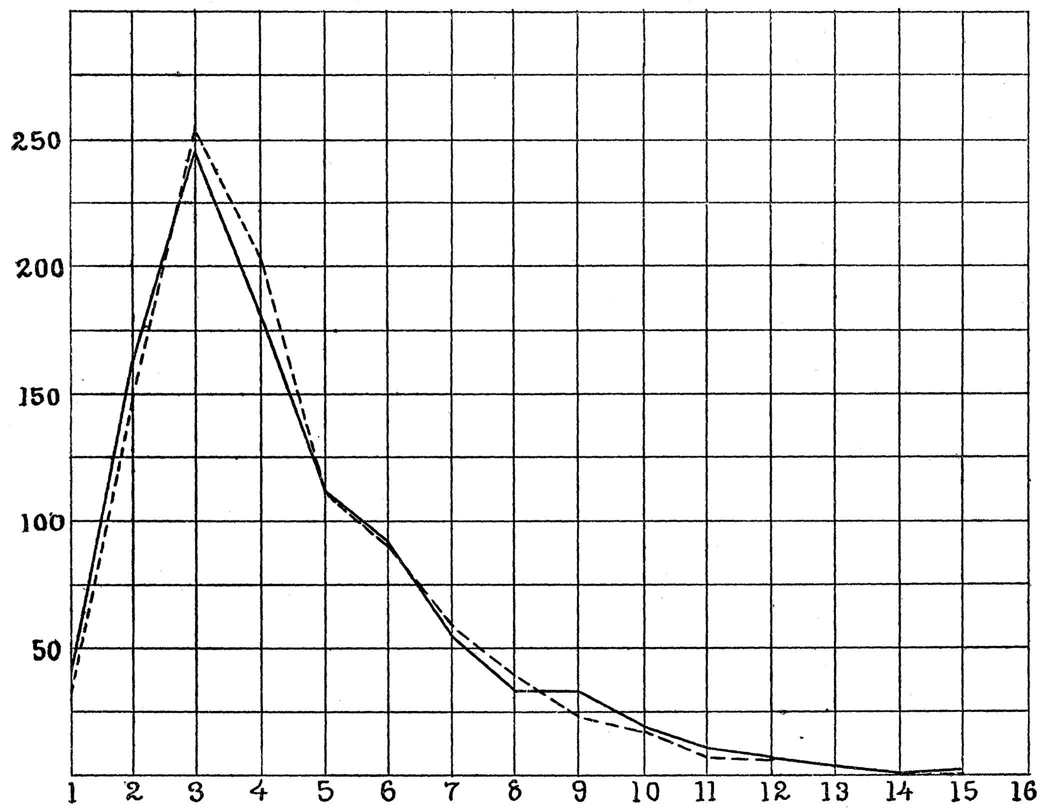

FIG. 4.—TWO GROUPS, OF FIVE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

The graphic representation of the results will be readily understood. It is only necessary to take a sheet of 'squared' paper, or paper ruled in two directions at right angles to each other, and, after placing the numbers showing letters in each word at points along a horizontal line separated from each other by equal distances, above each of these place other points whose distance from the base line shall be proportional to the number of such words in a thousand; then join these points by a broken line, and the characteristic curve is shown. Fig. 1 shows the curve thus constructed from the first thousand words in 'Oliver Twist,' the numerical analysis of which is shown above.

The next diagram (fig. 2) exhibits five curves constructed from the first five thousand words the same from work, in groups of one thousand each. It is presented in order to show the variation among groups based on a relatively small number of words.

The superiority of this method over that of simple word averages, as suggested by DeMorgan, is clearly shown in fig. 3, which exhibits two consecutive groups, of one thousand words each, from 'Vanity fair.' The numerical analysis of these groups is as follows: —

Letters................	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Words in 1st group	25	169	232	187	109	78	79	48	28	20	10	10	2	3
Words in 2d group	33	146	248	164	135	76	73	52	35	23	6	5	2	2

It will be seen that the total number of letters in the first group is 4,507, and in the second 4,508, or an average of 4.507 and 4.508 letters to each word in the respective groups. If this average, or 'mean word-length,' be alone considered, the two groups must be regarded as sensibly identical; but an inspection of the diagram shows that they are in reality quite different.

When the number of words in a group is increased to five thousand, the accidental irregularities begin to disappear, the curve becomes smoother, approximating more nearly to the normal curve which, it is assumed, is characteristic of the writer. Fig. 4 exhibits two groups, each of five thousand words, from 'Oliver Twist,' and it will be seen that considerable differences still exist. One of the curves shows an excess of nine-letter words, which does not appear in the other. They agree in showing a greater number of six-letter words than a smooth curve would demand. This excess may persist, and prove to be a real characteristic of Dickens's composition. Fig. 5 exhibits these two groups of five thousand words combined in one of ten thousand, giving a curve of greater smoothness, and approximating still more closely to the normal curve of the writer.

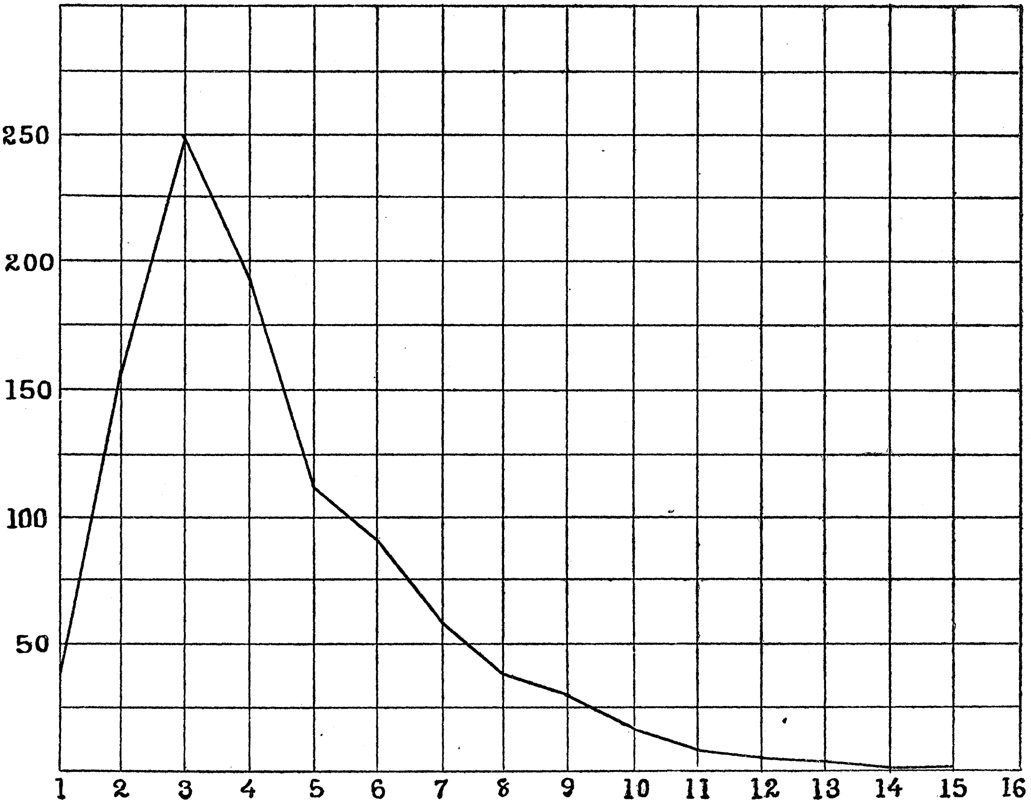

FIG. 5.—CURVE FOR TEN THOUSAND WORDS FROM 'OLIVER TWIST.'

In fig. 6, two groups of five thousand words each, from 'Vanity fair,' are shown; and in fig. 7, two groups of ten thousand each, from 'Oliver Twist' and 'Vanity fair,' are placed side by side for comparison, the former being represented by the continuous line, and the latter by the broken line. Although these curves differ, and while it is believed that the difference will persist with an increased number of words, it is certainly surprising, that in the analysis of ten thousand words from Dickens, and the same number from Thackeray, so close an agreement should be found. This agreement is particularly striking in words of eleven, twelve, and thirteen letters, the numerical comparison of which is as follows: —

Number of letters	11	12	13
Number of words in Dickens	85	57	29
Number of words in Thackeray	85	58	29

This closeness to identity must be largely the result of accident, and it would not be likely to repeat itself in another analysis.

The writer next examined was John Stuart Mill; and to test the persistence of form in compositions belonging to different periods of the author's life, and upon different subjects, two groups of five thousand words each were taken, — one from his 'Political economy,' and the other from his 'Essay on liberty.' It was anticipated, of course, that words of greater length would occur far more frequently than in the case of the novelists; but I confess to considerable surprise on finding from the very beginning, that although, on the whole, this anticipation was realized, the word which occurred most frequently was not the three-letter word, as with both Dickens and Thackeray, but the word of two letters. Indeed, the word of two letters was not only relatively more frequent, but absolutely; that is to say, it occurred more frequently in the composition of Mill than in that of either of the novelists, and with great uniformity, as it was in excess in each thousand of the ten analyzed. The explanation is easy, and is to be found in the liberal use of prepositions in sentence-building. The proposed method of analysis is designed to reveal any peculiarity of this kind, and the exemplification of its power thus early in the work was encouraging.

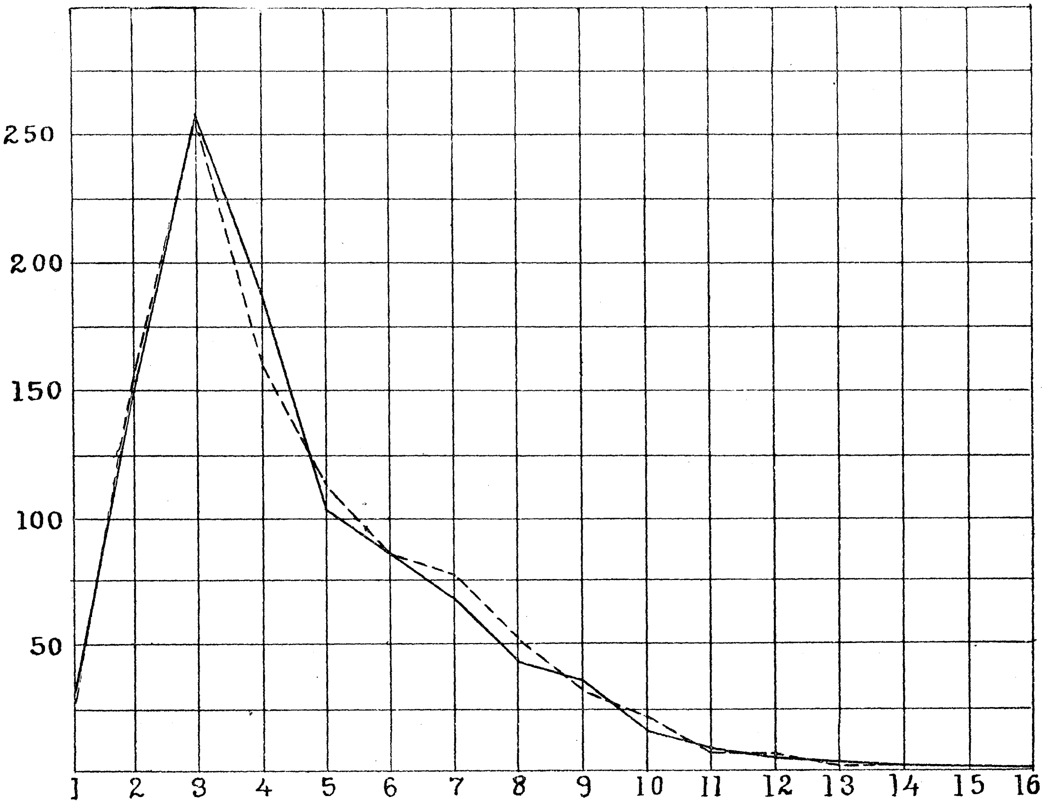

Fig. 6. — Two groups, of five thousand words each, from 'Vanity Fair.'

Figs. 8 and 9 show the curves for five thousand words from the 'Political economy' and from the 'Essay on liberty.' It will be observed, that, while they differ considerably, there is still, in a general way, a striking resemblance, and that they are in marked contrast with the curves of the novelists. An interesting case was furnished in two recent addresses on the labor question by Mr. Edward Atkinson. In reality, one address was given to two very different audiences. One was made up from the workingmen of Providence, and the other from the alumni of the Andover theological seminary. On reading the two, one cannot avoid being struck by the marked difference in style, although the two papers are much alike in substance. It was interesting, then, to inquire whether their curves of composition would show any marked resemblance. An analysis of five thousand words from each paper was made, and the result is shown in fig. 10. A very satisfactory, indeed a striking, general resemblance will be observed; and it will also be seen that Mr. Atkinson's curve differs decidedly from others previously figured and described. It is shown in contrast with that of John Stuart Mill in fig. 11. Mr. Atkinson's composition is remarkable in respect to the shortness of the words used. The average length of ten thousand words in his addresses on the labor question is 4.298 letters. The mean word-length of the writers thus far examined, based upon a count of ten thousand words from each, is as follows: —

Atkinson	4.298
Dickens	4.342
Thackeray	4.481
Mill	4.775

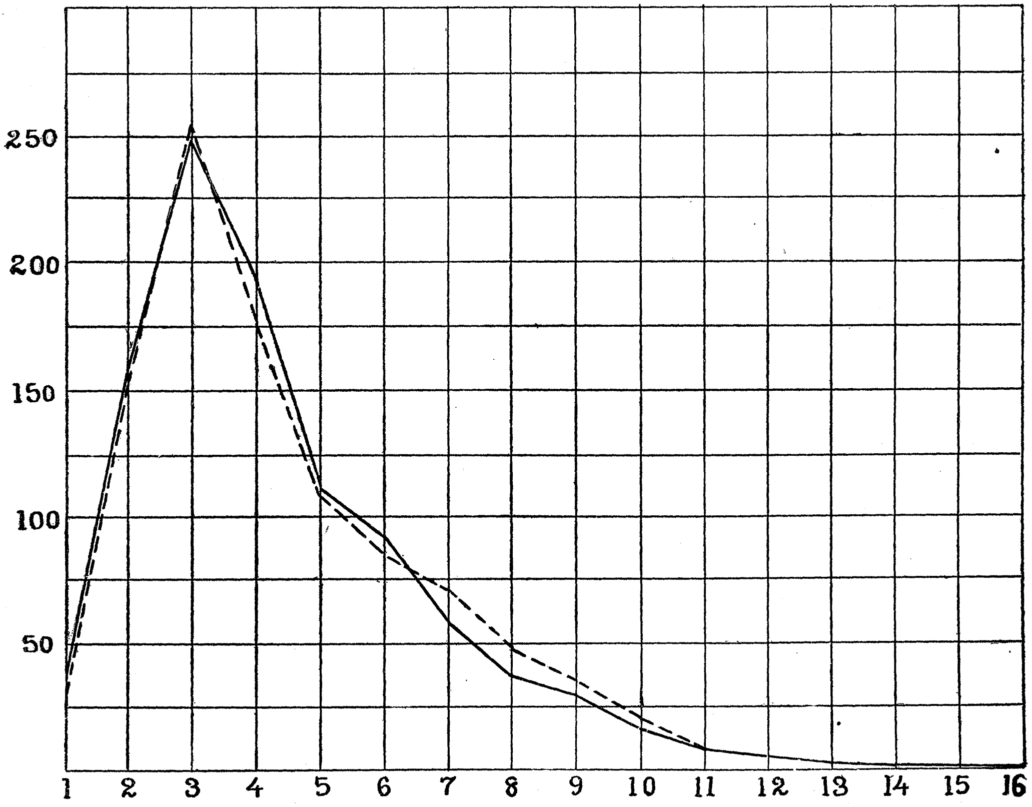

FIG. 7.—TWO GROUPS, OF TEN THOUSAND WORDS EACH, FROM 'OLIVER TWIST,' ———; AND FROM 'VANITY FAIR,' — — — —.

A friend has furnished me with the result of the count of the first five thousand five hundred words of Caesar's 'Commentaries.' The mean word-length is 6.065. The most extensive word-counting that I know of is that of the words and letters in the Bible. I cannot vouch for the reliability of the information which periodically floats through the columns of the public press, that the Old Testament contains 592,493 words with 2,728,100 letters, and the New Testament 181,253 words with 838,380 letters. It is interesting to note, however, that these numbers give averages of 4.604 and 4.625 respectively, agreeing within less than one-half of one per cent.

Before making an analysis of Mr. Atkinson's composition, and after having counted more than thirty thousand from other writers, I had concluded that a group of one thousand words whose average length was less than four letters would not occur, except in compositions especially written in short words. Out of ten such groups from Mr. Atkinson's addresses, however, one was found whose mean word-length was 3.991. I have recently received from him a brief paper, entitled 'How do we all get a living?' which was published in *Work and wages*, and in the preparation of which he made a special effort to use the simplest language possible. The article contains a little more than two thousand words, the number being too small for the construction of a curve which would be comparable with those already exhibited. The general form of one based upon two thousand words is similar to that previously obtained from the same writer, and the mean word-length is 3.771.

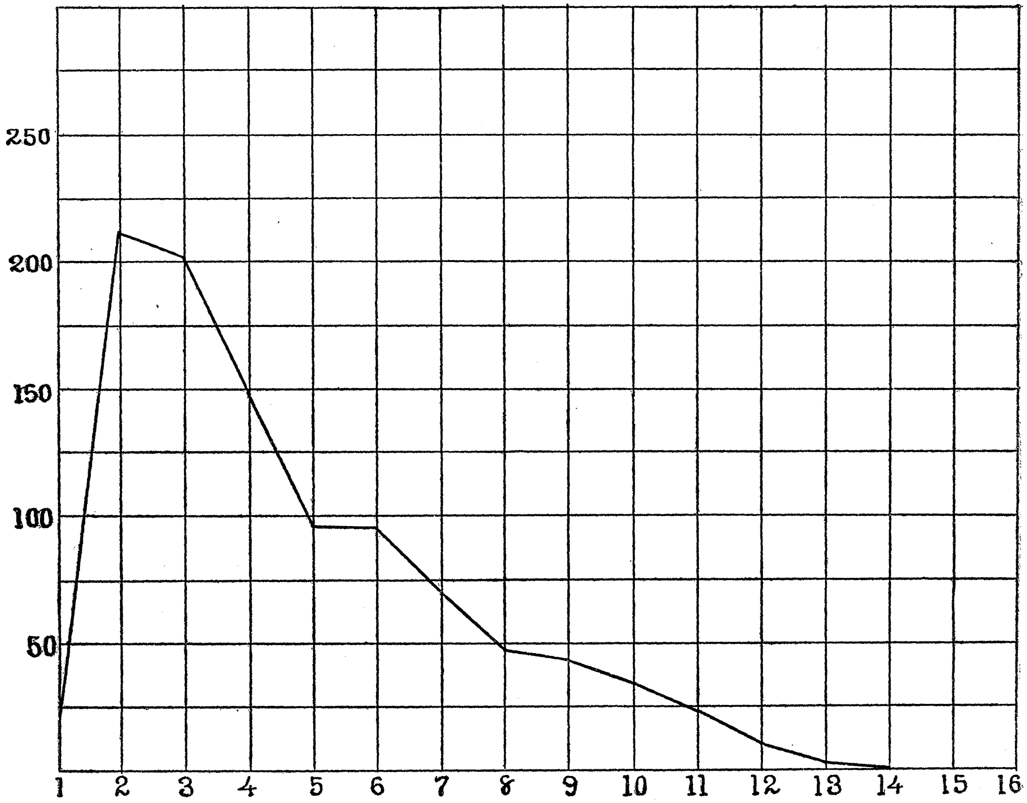

FIG. 8. — CURVE OF FIVE THOUSAND WORDS FROM MILL'S 'POLITICAL ECONOMY.'

Interesting evidence of the validity of this method of analysis and identification has been furnished by several friends who have had the patience to enumerate the letters in many thousand words from different sources. Prof. Stanley Coulter sends me the result of a count of ten thousand from Dickens's 'Christmas carol.' He writes, "I became exceedingly interested in watching how little tricks of composition affected the 'curve.' For instance, one of the characters, 'Scrooge,' appears in one place very often, and an excess of 7's is the result; in another place 'Fizziwig,' and the 8's creep up [this is doubtless owing to the frequent appearance of the names]. Other variations and excesses seem to come from Dickens's love of certain forms of description, which he iterates and reiterates upon a single page."

I have plotted these ten thousand words from the 'Carol' with the ten thousand already shown from 'Oliver Twist,' in fig. 12. A very close resemblance will be observed, and it will be noticed that the *mean* of these two curves would be free from certain irregularities which occur in both, and would be a much closer approximation to the normal characteristic curve of Dickens.

It is hardly necessary to say that the method is not necessarily confined to the analysis of a composition by means of its mean word-length: it may equally well be applied to the study of syllables, of words in sentences, and in various other ways. The results thus far obtained from its application would appear to justify the claim that it is worthy of a thorough test through which the validity of its assumptions might be proved or disproved. Its principal merits are, that it offers a means of investigating and displaying the mere mechanism of composition, and that it is purely mechanical in its application. In virtue of the first, it might reveal characteristics which a writer would make no attempt to conceal, being himself unaware of their existence; and, of the second, the conclusions reached through its use would be independent of personal bias, the work of one person in the study of an author being at once comparable with that of any other.

Many interesting applications of the process will suggest themselves to every reader; the most notable, of course, being the attempt to solve questions of disputed authorship, such as exist in reference to the letters of Junius, the plays of Shakspeare, and other less widely known examples. It might also be utilized in comparative language studies, in tracing the growth of a language, in studying the growth of the vocabulary from childhood to manhood, and in other directions too numerous to be catalogued. An illustration of its application to another language is shown in the analysis of more than five thousand words in Caesar's 'Commentaries,' already referred to, which is represented in fig. 13. The curve shows a relatively large use of long words, and its peculiar feature is the evident indication of two maximum ordinates nearly equal to each other.

From the examinations thus far made, I am convinced that one hundred thousand words will be necessary and sufficient to furnish the charac-

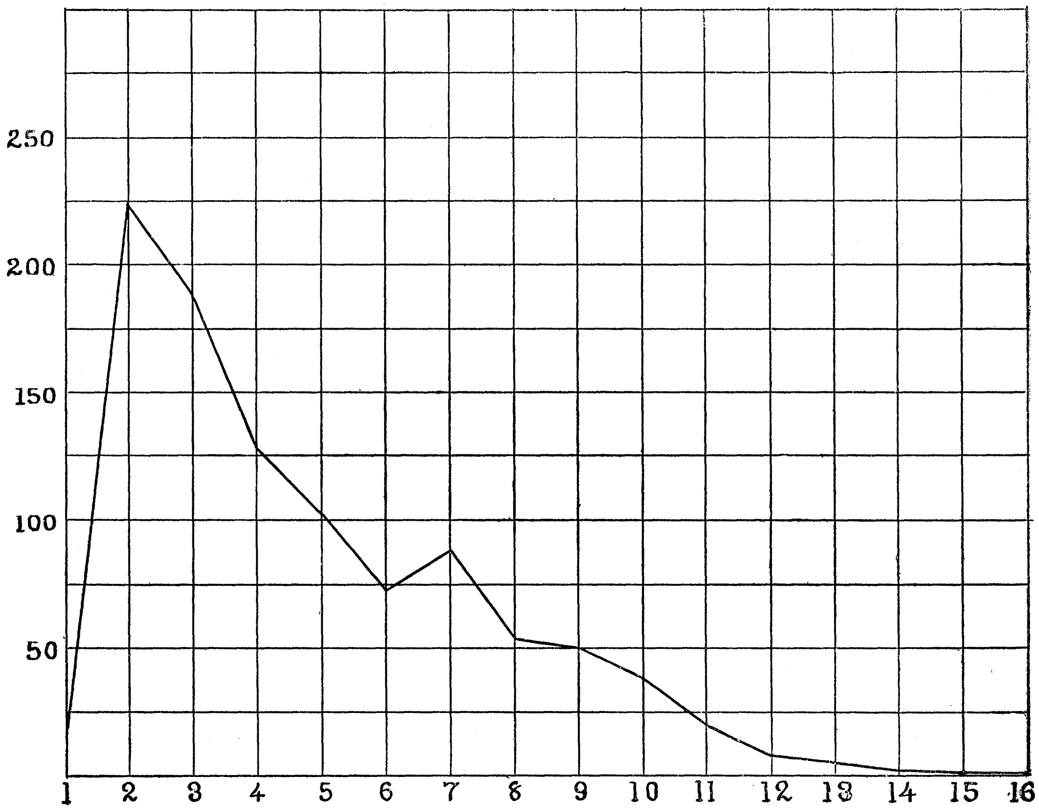

FIG. 9.—CURVE OF FIVE THOUSAND WORDS FROM MILL'S 'ESSAY ON LIBERTY.'

teristic curve of a writer, — that is to say, if a curve is constructed from one hundred thousand words of a writer, taken from any one of his productions, then a second curve constructed from another hundred thousand words would be practically identical with the first, — and that this curve would, in general, differ from that formed in the same way from the composition of another writer, to such an extent that one could always be distinguished from the other. To demonstrate the existence of such a curve will require the enumeration of the letters in several hundred thousand words from each of a number of writers. Should its existence be established, the method might then be applied to cases of disputed authorship. If striking differences are found between the curves of known and suspected compositions of any writer, the evidence against identity of authorship would be quite conclusive. If the two compositions should produce curves which are practically identical, the proof of a common origin would be less convincing; for it is possible, although not probable, that two writers might show identical characteristic curves.

T. C. Mendenhall.

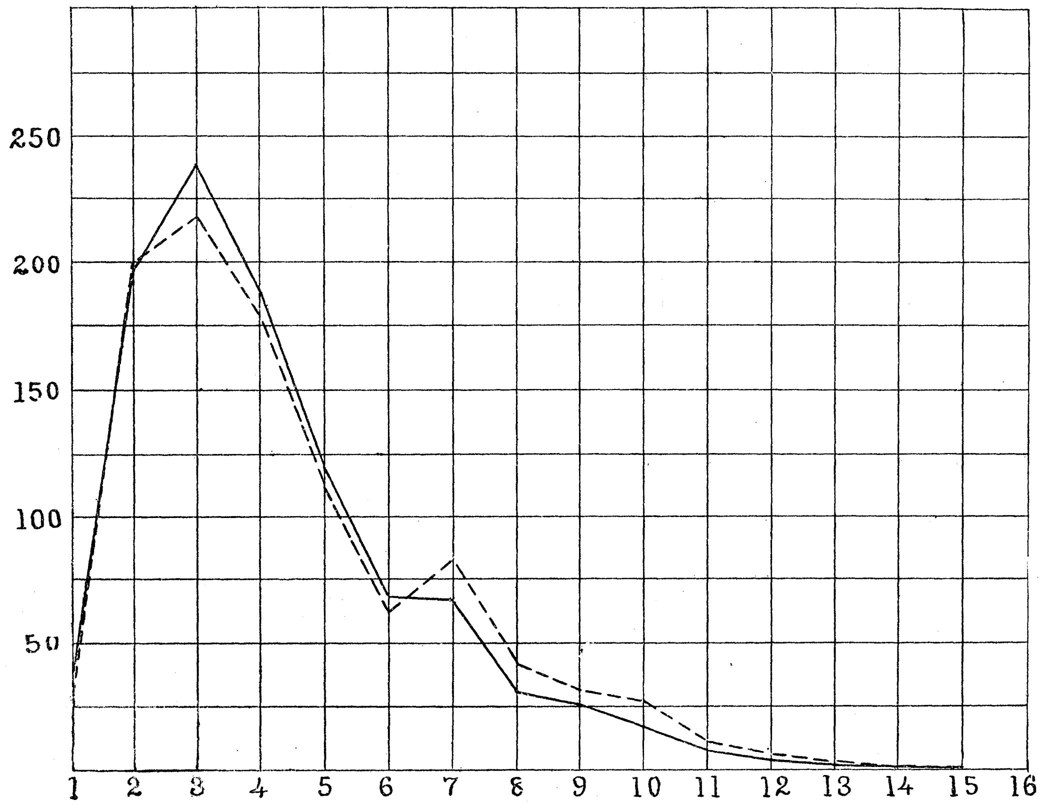

Fig. 10. — Two groups, of five thousand words each, from addresses of Edward Atkinson: Address to workingmen, ———; to alumni of theological seminary, — — — —.

TIDAL OBSERVATIONS OF THE GREELY EXPEDITION.

The principal tidal observations were made at Fort Conger, on Lady Franklin Bay, by various members of the expeditionary force working under direction of Sergt. Edward Israel, and with a general supervision by the commanding officer of the expedition. They consisted of hourly heights of the tide from Aug. 20, 1881, to July 1, 1882, and the times and heights of high and low waters from Aug. 20, 1881, to June 30, 1883, both series read from fixed staff gauges and practically continuous. A broken series of high and low waters from July 1 to Aug. 8, 1883, obtained under unfavorable conditions, were not used in the discussion. There were also short series at seven outlying stations on the coasts of Greenland and

Grinnell Land, and a casual observation of high water at the head of Greely Fiord, during the progress of the readings at Fort Conger, with a dozen or more high and low waters noted during the retreat through Kennedy Channel and Kane Basin. The original records, too bulky for easy transportation, were left stored at Fort Conger when the party abandoned that station; but close transcripts, previously prepared and carefully verified, were brought away by Lieutenant Greely, and on his return to this country referred to the superintendent of the coast and geodetic survey, and later were placed in the hands of Mr. Alex. S. Christie, chief of the tidal division of the office, for reduction and discussion.

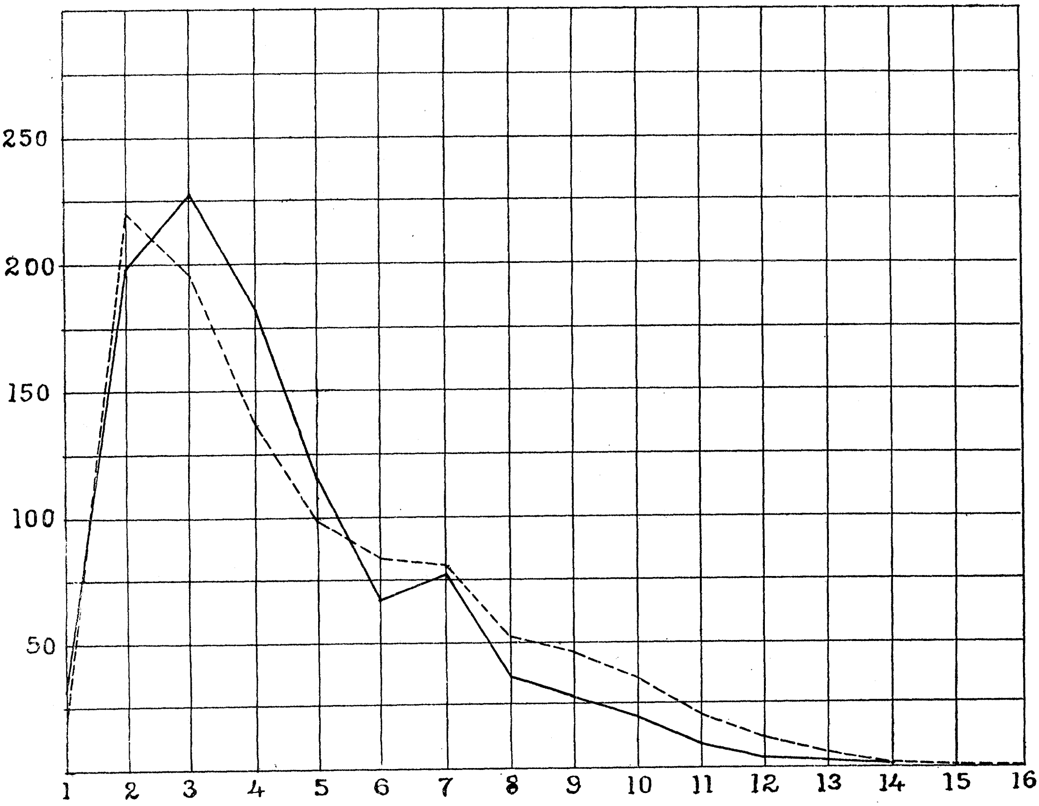

FIG. 11.—TWO GROUPS, OF TEN THOUSAND WORDS EACH. ATKINSON, ———: MILL, — — — —.

The weak point of tidal observations is almost invariably, even in middle latitudes, the instability of the staff and the undetermined fluctuations in altitude of the staff zero; so that it not infrequently happens that a satisfactory reduction of all the observations to the same plane of reference is a wholly intractable problem. In high latitudes, at stations beset with heavy ice, even short series are, as a rule, sadly out of joint and comparatively worthless. Unless the stability of the gauge is absolutely assured, which can seldom be the case, only frequent resort to levels between the gauge and one or more permanent bench-marks on shore can insure scientific value to the observations. At Fort Conger the observations of the first year depended in this respect upon a gauge that seems to have been stable, those of the second upon series of spirit-levels. Only two months of the series were in much doubt, and a tolerably satisfactory adjustment of these was finally effected. The observations bear abundant internal evidence of a conscientious and persistent endeavor to secure trustworthy and precise results; and, although they are far from equalling observations of standard excellence in middle latitudes, they are believed to constitute the longest and best series ever brought back from the arctic seas.

Following are some of the results of a non-harmonic analysis of the observations at Fort Conger:

the mean lunitidal intervals are $11^h\ 33^m.3$ and $17^h\ 45^m.3$; the mean range is 1.328 metres; the semi-mensual curves for intervals and heights give the age of the tide 1.4 days, the moon tide 2.2 times the sun tide, and satisfy closely the equilibrium formulae of Bernouilli. The diurnal inequality in height is, in comparison with the whole tide, three times as small as in Smith Sound; the influence of the sun in producing it is practically equal to that of the moon; it vanishes for high water $2^d\ 22^h$ after, for low water $1^d\,08^h$ before, the vanishing of the moon's declination, and the interval of the two former events appears to be independent of the solar declination. A method of graphical analysis, due to the late Assistant L. F. Pourtales of the U.S. coast survey, brings out the fact that the diurnal inequality at Fort Conger is caused by a wave that has a sidereal day for its mean period; the same thing is also shown to obtain at Port Foulke and Van Rensselaer harbor in Smith Sound, and at Thank God harbor in Polaris Bay. The general phenomenon of diurnal inequality is then taken up as a problem in kinematics, the diurnal inequality wave is analyzed into its principal components, and the sidereal period shown to have place at still other stations both within and without the arctic circle, and to be a rule rather than an exception. The results of an harmonic analysis of the first year's observations will be found in the report: in so far as they relate to the same matters, they confirm the results previously found and stated above.

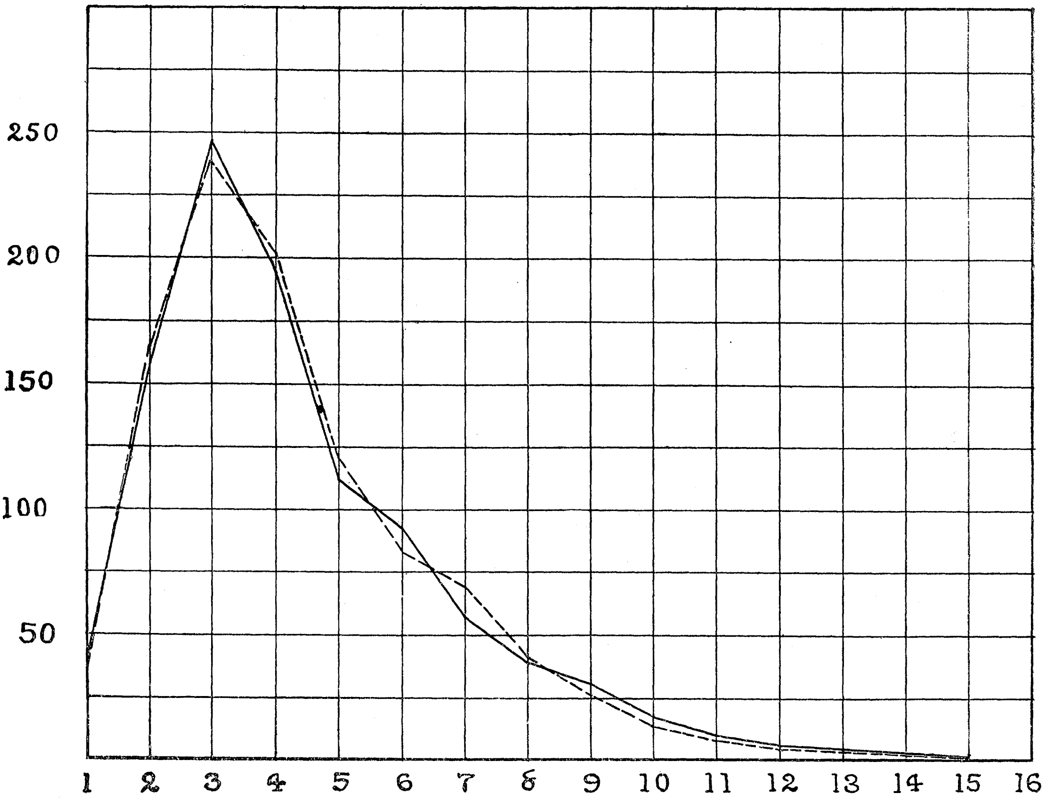

Fig. 12.—Two groups, of ten thousand words each, from Dickens: 'Oliver Twist,' ——; 'Christmas Carol,' — — — —.

A comparative study of the specific characters of the Fort Conger and other arctic tides with the cotidal lines, widths, and depths of the tidal avenues to the Polar Ocean, with whatever other tidal data from high latitudes was accessible, resulted in certain inferences stated in the report, and which may perhaps be tolerated here. The weakness of the tide-producing forces near the pole and a propensity to dissipate as a free wave as soon as formed, in waters of even moderate depth, are two causes operating to prevent the generation of local tides of appreciable magnitude in that region. The tides of the Pacific are not likely to make themselves felt in that vast expanse through a strait only some forty miles in width and less than thirty fathoms in depth, with far-stretching shoal approaches on either side. On the other hand, the relation of the Polar to the Atlantic Ocean is so intimate as to amount to identity. The continuity of the Atlantic basin has been demonstrated by soundings up to and beyond the 80th parallel. The channel between Spitzbergen and the European coast is about a hundred fathoms deep and four hundred miles in width; that between Spitzbergen and Greenland has about the same width, but is one, two, and three miles deep. The tides of the circumpolar seas cannot avoid forming a part of the Atlantic system. As to the tide in Lady Franklin Bay, it seems almost a certainty that it is chiefly an Atlantic tide that has flowed up through the Spitzbergen Sea, rounded Greenland, and entered Robeson Channel from the north, where it probably meets another and fainter Atlantic tide from the south, which, delayed and spent in the shallow West Greenland seas, comes into Lady Franklin Bay two or three hours later.

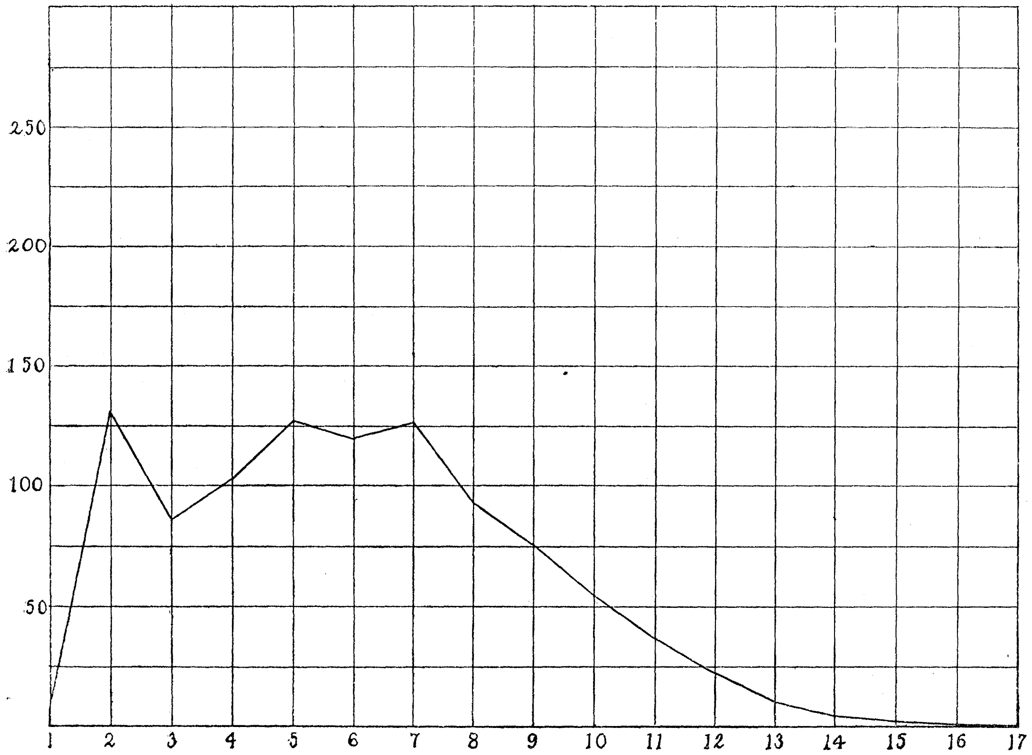

FIG. 13.—GROUP OF FIVE THOUSAND FIVE HUNDRED WORDS FROM CAESAR'S 'COMMENTARIES.'

But the laws of the tides in the circumpolar seas, a *cul de sac* into which run the tides of an ocean stretching from pole to pole, and where the absence of controlling astronomical forces is favorable to tidal anarchy, can only be determined with certainty from long series of observations at stations generously distributed about the polar basin. The establishment and maintenance by Lieutenant Greely of one such station, and his preservation of the records of observation, will be regarded as substantial services to science by all interested in this branch of physical inquiry.

A. S. C.

AGRICULTURE IN ENGLAND IN 1886.

IN outlining, in a recent number of *Science* (ix. No. 212), the reports presented by the British commission on the existing trade depression, special attention was called to the fact that it was admitted on all hands that the agricultural classes were the worst sufferers. The lower prices of agricultural produce were very far-reaching in their consequences. For this reason the latest returns as to that produce are of timely interest; and we con-